SUPPORTING INFORMATION of From residue co-evolution to protein conformational ensembles and functional dynamics

Ludovico Sutto, Simone Marsili, Alfonso Valencia, and Francesco L. Gervasio

I. PAIRWISE MODEL INFERENCE

As discussed in the main text, in this work we build statistical models for distributions over sequences of evolutionary-related proteins, arranged in a multiple sequence alignment (MSA). A generic MSA consists in a $M \times L$ matrix having M sequences as rows:

$$\boldsymbol{A} \equiv \begin{pmatrix} \boldsymbol{a}^1 \\ \boldsymbol{a}^2 \\ .. \\ \boldsymbol{a}^M \end{pmatrix} \tag{1}$$

where each $\boldsymbol{a}^s$ is an array with L entries from an alphabet of q=21 letters (the 20 natural amino acids plus the gap state):

$$\boldsymbol{a}^s \equiv \begin{pmatrix} a_1^s & a_2^s & .. & a_L^s \end{pmatrix}. \tag{2}$$

In particular, we consider maximum entropy, pairwise models for protein families[1–3], which reproduce the amino acid frequencies for single residues and correlations between amino acids of different residues as observed in the MSA. Let $F_i(\alpha)$ denote the frequency of amino acid α at position i and $F_{i,j}(\alpha,\beta)$ the frequency of co-occurence of amino acids α and β at positions i and j, respectively. In the following, we will use uppercase F for averages over the MSA sequences, lowercase f for averages over the model distribution and θ for model parameters. In order to reduce the effect of possible sampling biases in the MSA, for each sequence s we computed a weight $w_s = 1/n_s$, where n_s is the number of similar sequences to sequence s, using a similarity threshold of 0.7. Then, the frequencies $\boldsymbol{F} \equiv \{F_i(\alpha)\}, \{F_{i,j}(\alpha,\beta)\}$ were calculated as weighted averages over the M sequences in the MSA[2, 3]:

$$\begin{aligned} F_i(\alpha) &= M_{\text{eff}}^{-1} \sum_{s=1}^{M} w_s \delta(a_i^s, \alpha) \\ F_{i,j}(\alpha,\beta) &= M_{\text{eff}}^{-1} \sum_{s=1}^{M} w_s \delta(a_i^s, \alpha)\delta(a_j^s, \beta) \end{aligned} \tag{3}$$

where $M_{\text{eff}} = \sum_s w_s$ and $\delta(..)$ returns 1 if the arguments are equal and 0 otherwise. On using a set of Lagrange multipliers $\boldsymbol{\theta} \equiv \{h_i(\alpha)\}, \{J_{i,j}(\alpha,\beta)\}$ to constrain the model averages $\boldsymbol{f} \equiv \{f_i(\alpha)\}, \{f_{i,j}(\alpha,\beta)\}$ to the observed frequencies $\boldsymbol{F}$, the maximum entropy distribution takes the form:

$$P(\boldsymbol{a}) = Z(\boldsymbol{\theta}^*)^{-1} \exp \underbrace{\left(\sum_i h_i^*(a_i) + \sum_{i,j>i} J_{i,j}^*(a_i, a_j) \right)}_{-H(\boldsymbol{a})} \tag{4}$$

where $H(\boldsymbol{a})$ is the energy of sequence $\boldsymbol{a}$, $Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{a}} \exp[-H(\boldsymbol{a})]$ is the partition function, and the parameters $\boldsymbol{\theta}^*$ satisfy the equations:

$$f_p^* = \left(\frac{\partial \ln Z}{\partial \theta_p} \right)_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = F_p, \qquad p = 1, \binom{L}{2} q^2 + Lq. \tag{5}$$

In practice, to solve Eqs. 5 we minimized the following quantity:

$$L(\boldsymbol{\theta}) = \ln Z(\boldsymbol{\theta}) - \sum_p F_p \theta_p + (\lambda/2) \sum_p \theta_p^2 \tag{6}$$

that corresponds to the negative log-likelihood of the parameters $\boldsymbol{\theta}$ (rescaled by M) plus a l_2-regularization term. As observed in [4], even though the original problem is overparametrized, due to this latter term $L(\boldsymbol{\theta})$ has a unique minimum $\boldsymbol{\theta}^*$. In principle, this minimum could be found iteratively, computing the gradient of $L(\boldsymbol{\theta})$ at every iteration:

$$\left(\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_p} \right)_{\boldsymbol{\theta}^t} = f_p(\boldsymbol{\theta}^t) - F_p + \lambda \theta_p^t \tag{7}$$

and updating $\boldsymbol{\theta}$ proportionally:

$$\theta_p^{t+1} = \theta_p^t - \alpha \left(\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_p} \right)_{\boldsymbol{\theta}^t} \tag{8}$$

where the parameter α represents a step-size. At each iteration t, the model averages $\boldsymbol{f}$ at $\boldsymbol{\theta}^t$ (and the gradient through Eq. 7) were estimated by multiple Monte Carlo simulations (from 20 to 64) and a number of sweeps (moves per residue in the chain) ranging from 10^4 to 10^5. Since convergence of gradient descent (Eq. 8) was prohibitively slow, we adopted an accelerated gradient iterative method with an additional momentum step[5]:

$$\begin{aligned} &1: \quad \theta_p^{t+1} = \eta_p^t - \alpha \left(\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_p} \right)_{\boldsymbol{\eta}^t} \\ &2: \quad \eta_p^{t+1} = \theta_p^{t+1} + \beta^t (\theta_p^{t+1} - \theta_p^t) \end{aligned} \tag{9}$$

starting from $\boldsymbol{\eta}^0 = \boldsymbol{\theta}^0 = 0$. The step-size α has been kept fixed during the minimization, using a constant value of $\alpha = 0.01$. The value of momentum-related parameter

β^t depends on the iteration number as $\beta^t = \frac{t}{t+3}$. After 2000 iterations of accelerated gradient descent, momentum was turned off by setting $\beta = 0$ and parameters were refined with 3000 iterations of standard gradient descent. Finally (see Results section in the main text), we checked through extensive sampling (10^6 sweeps per trajectory) that the inferred parameters do correspond to the desired minimum and that they reproduce the amino acid frequencies from the MSA, $\boldsymbol{f}^* \approx \boldsymbol{F} - \lambda\boldsymbol{\theta}^*$ (see Eq. 7). In the left panel of Fig. S1 the reconstructed pairwise frequencies $f_{i,j}(\alpha,\beta)$ are compared with the frequencies $F_{i,j}(\alpha,\beta)$ computed from the corresponding MSA for two of the protein domains in the dataset, the small RRM_1 domain and the largest domain in the set, the Trypsin domain.

II. COMPUTATION OF CO-EVOLUTIONARY COUPLINGS AND CONTACT PREDICTION

Co-evolutionary couplings were extracted from the converged values of the parameters $\boldsymbol{J}$ following a protocol proposed by Ekeberg et al.[4]: for each pair of residues i and j, we first computed the Frobenius norm of the corresponding $q \times q$ couplings sub-matrix:

$$c_{i,j} = \|\boldsymbol{J}'_{i,j}\|_2 = \left(\sum_{\alpha=1}^{q}\sum_{\beta=1}^{q} J'_{i,j}(\alpha,\beta)^2\right)^{1/2} \tag{10}$$

where $J'_{i,j}$ denotes the $q \times q$ sub-matrix $J_{i,j}$ after double-centering. Then, we applied an average product correction[6]:

$$C_{i,j} = c_{i,j} - \langle c\rangle_i \langle c\rangle_j / \langle c\rangle \tag{11}$$

where the $\langle c\rangle$ denotes an average over all the elements of c, and $\langle c\rangle_i$ and $\langle c\rangle_j$ averages over the i-th and j-th columns, respectively. In the right panel of Fig. S1, the C_β-C_β distance distributions corresponding to set of pairs of residues with different values of co-evolutionary couplings are compared. We included in the analysis all the residues pairs separated by more than 4 residues along the sequence, and averaged across the 18 protein families. Positions in the sequence alignments with a fraction of gaps $> 5\%$ were discarded in order to minimize the effect of misaligned regions. Strongly coupled residues ($C > 0.7$, red line) are separated by short distances distributed around a mean value of $\sigma = 5.5$ Å, with a spread of ~ 2 Å, in agreement with the position of the first peak in the pair correlation function characteristic of the nearest neighbors of a central residue, as computed from the same dataset of protein structures (black broken line).

The top N ranked pairs of residues according to the coupling $C_{i,j}$ for $j > i + 4$, with N being the number of amino acids in a protein structure, were classified as "long-range" predicted contacts. Furthermore, we defined a linear combination of scores as a heuristic for determining if two residues $i, i+4$ are in contact within an α-helix structure:

$$\begin{aligned} C_\alpha(i,i+4) = &\frac{1}{3}\sum_{k=-1}^{1} C(i+k, i+4+k) - \\ &\frac{1}{5}\sum_{k=-1}^{3} C(i+k, i+2+k) \end{aligned} \tag{12}$$

This function quantifies the consistency with the helical geometry of the couplings involving residues close in sequence to i and $i + 4$, and tends to be positive for $i \to i + 4$ pairs within a α-helix. Pairs having a value of $C_\alpha > 0.05$ were predicted as contacts in helices. Fig. S2 compares the actual and predicted alpha-helical content per residues of the 18 domains analyzed in this work.

III. COMPARISON WITH MEAN FIELD ALGORITHM

Fig. S3 shows the precision of the top ranked predictions as a function of their scaled rank, for the 18 protein domains analyzed in this work. For a structure with a total of NC contacts, the precision for scaled rank r is the fraction of the $r \times$NC top-scoring pairs of residues that are actually found in contact (C_β-$C_\beta < 8$ Å). We included in the analysis all the residue pairs with a sequence separation larger than four (dark blue lines) and larger than five (light blue lines). Fig. S3 shows also (broken lines) the precision curves obtained following exactly the same protocol (and the same definition of score, Eqs. 10 and 11) but using the mean field approach introduced in [3]. Following [3], overfitting was avoided adding a pseudocount $\lambda = M_{\text{eff}}$ in the calculation of the empirical covariance. Even if the number of cases taken into consideration (18 protein families) is too small to make a quantitative comparison between the two methods, this analysis shows that predictions from Boltzmann learning have better or similar precision to the predictions obtained using a mean field approximation. In some cases these improvements are quite evident (*e.g.* the HTH_31, Sigma70_r2 and RRM_1 domains).

IV. LIKELIHOOD GRADIENT ESTIMATION AND MULTIMODAL DISTRIBUTIONS OF SEQUENCES

The potential organization of the MSA in clusters, representing families in a superfamily or subfamilies of sequences could represent a problem for our importance-sampling strategy. In particular, the number of MCMC steps at each gradient evaluation (see Eq. 7) should be large enough for the simulated trajectories to equilibrate over different "basins" in sequence space, minimizing the error in the numerical estimates of the gradient of the likelihood in Eq. 6. In the main text

(Fig. 1, panels A and B), we show that this is true for the ADH_zinc_N protein family: our simulated trajectories reproduce almost quantitatively the empirical distribution of sequences $P(\mathrm{PC1}, \mathrm{PC2})$ over the first two PCA eigenvectors of the MSA covariance matrix[7]. In particular, the "landscape" associated to this distribution - that is given to within an additive constant by $A(\mathrm{PC1}, \mathrm{PC2}) = -\log P(\mathrm{PC1}, \mathrm{PC2})$ - nicely characterizes the complexity of the system in terms of minima (metastable clusters of sequences) and the barriers separating them. In Fig. 4 we show four examples of different complexity: for the two protein families RRM_1 and cNMP_binding, the distribution is essentially unimodal (top panels), while the two families TrkA_N and CMD clearly show multiple minima separated by barriers large enough to slow down the MCMC sampling. We roughly estimated the characteristic time-scale of global equilibration from the simulated trajectories as the exponential autocorrelation time of the slowest relevant component of the energy $H(\boldsymbol{a})$. The fitted τ_{slow} are reported in table I, in units of MC sweeps (moves per residue). As expected, families associated with a multimodal distribution of sequences are characterized by a slow process with a relaxation time of $\approx 10^3$ MC sweeps, that is clearly absent for the families associated to unimodal distributions. Albeit the actual values depend on the specific details of the MCMC dynamics, the observed effect is general and clearly system-dependent. In order to check the effect of these different time-scales on the reconstruction errors, we analyzed again the protein families in table I, using 16 simulated trajectories of length 10 τ_{slow} sweeps at each minimization step. In this way, the accuracy of gradient estimation is tuned to the system-dependent slowest relaxation time. The final mean relative absolute errors $\epsilon_r = \langle |F_{i,j} - f_{i,j}|/F_{i,j}\rangle_{F_{i,j}>0.01}$ are similar for the different systems and lie in the range of 1.5-3.5, demonstrating that the optimal number of MCMC steps should be tailored to the time-scale associated to global equilibration. Using an in-house Fortran implementation of our Boltzmann learning approach, a single gradient evaluation on 10^4 sweeps required 1 second for a domain of 60 residues, 3 seconds for a domain of 100 residues, and 28 seconds for the largest domain in the set, the Trypsin domain (263 residues). By using 5000 total minimization steps and 16 independent MC simulations per gradient evaluation, this translates to a total computational cost for full convergence of 1, 3 and 622 hours, respectively.

V. PROTEIN COARSE-GRAINING AND HAMILTONIAN

The protein chain is described by the heavy atoms of the backbone plus the C_β atoms of the side chains. The potential V acting on the protein can be written as the sum of a bonded term V_b and a non-bonded term V_{nb} as $V = V_b + V_{nb}$ where:

$$V_b = \widetilde{V}_{bonds} + \widetilde{V}_{angles} + \widetilde{V}_{torsions} + V_{\phi,\psi} \quad (13)$$

$$V_{nb} = V_{LJ}^{C_\beta C_\beta} + V_{LJ}^{O..N} + \widetilde{V}_{LJ}^{1-4} + V_H \quad (14)$$

The potential contributions taken from the AMBER99SB-ILDN force field [13] are denoted by a tilde. These terms account for the correct backbone geometry and comprise the two-body bonded potential $\widetilde{V}_{bonds}$, the three-body angle potential $\widetilde{V}_{angles}$, the four-body proper (but the ϕ and ψ angles) and improper dihedrals $\widetilde{V}_{torsions}$ and the $\widetilde{V}_{LJ}^{1-4}$ short-range van der Waals potentials (two-body term between atoms separated by three bonds). Their precise functional forms and constants are fully described elsewhere[14]. On top of the absence of explicit solvent, the present model also lacks the hydrogen atoms and, most importantly, the atomic charges. For these reasons, it is crucial to recalibrate some of the force field terms to effectively take these effects into account in the framework of the current coarse-graining. To this aim, we re-fitted the force field proper dihedral potentials for the ϕ and ψ Ramachandran's angles resulting in the $V_{\phi,\psi}$ term (see Fig. S10 and Table S1). Finally, the remaining non-bonded terms account respectively for the co-evolution prediction $V_{LJ}^{C_\beta C_\beta}$, for an hydrogen bond mimicking potential for helix stabilization $V_{LJ}^{O..N}$ and for a generic hardcore repulsion preventing atom overlaps V_H. These terms can be explicitly written as the following sums:

$$V_{LJ}^{C_\beta C_\beta} = \sum_{i,j \in \mathcal{I}} V_{LJ}^{r_{\beta\beta},\epsilon_{\beta\beta}}(r_{ij}) \quad (15)$$

$$V_{LJ}^{O..N} = \sum_{a \in \mathcal{A}} V_{LJ}^{r_{ON},\epsilon_{ON}}(r_{O_a,N_{a+4}}) \quad (16)$$

$$V_H = \sum_{i,j \notin \mathcal{I}} \left(\frac{r_H^{\beta\beta}}{r_{ij}}\right)^{12} + \sum_{\substack{i=C_\beta \\ j \neq C_\beta}} \left(\frac{r_H^{\beta x}}{r_{ij}}\right)^{12} + \sum_{\substack{i \neq C_\beta \\ j \neq C_\beta}} \left(\frac{r_H^{xx}}{r_{ij}}\right)^{12} \quad (17)$$

Where N is the number of residues of the protein, $\mathcal{I}$ is the set of the N best predicted pair interactions between C_β atoms distant at least 5 residues along the sequence, r_{ij} is the distance between atom i and j, $\mathcal{A}$ is the set of predicted α helical residues, $r_{O_a,N_{a+4}}$ is the distance between the oxygen atom of residue a and the nitrogen atom of residue $a+4$, $V_{LJ}^{r_0,\epsilon_0}(r) = \epsilon_0 \left[(r_0/r)^{12} - 2(r_0/r)^6\right]$ is the usual 12-6 Lennard-Jones function and the remaining parameters are set to: $r_{\beta\beta} = 0.55$ nm, $\epsilon_{\beta\beta} = \epsilon_{ON} = 15$ kJ/mol, $r_{ON} = 0.3$ nm, $r_H^{\beta\beta} = 0.5$ nm, $r_H^{\beta x} = 0.3$ nm, $r_H^{xx} = 0.2$ nm. See the inset in Fig. S5 for a schematic illustration of all the non-bonded potentials and the protein coarse-graining.

VI. FOLDING SIMULATION DETAILS

For all the 18 proteins, we used the same simulation protocol consisting of a Langevin dynamics simulation performed using the molecular dynamics package GROMACS 4[15], with an inverse friction constant of 1 ps and an integration time step of 2 fs. Each simulation was set up with a replica exchange scheme in order to enhance the sampling, with 10 replicas at increasing temperatures (230 K, 245 K, 260 K, 276 K, 294 K, 313 K, 333 K, 354 K, 377 K, 401 K) allowed to attempt an exchange every 40 ps. Given the small number of simulated atoms with respect to an all-atom simulation, the energy fluctuations were still large enough for the largest protein to lead an average exchange rate of 10% between adjacent replicas with the above temperature scheme. The starting conformation of each folding run is an unfolded conformation obtained after 2 ns at high temperature with all the predicted interactions turned off. This led to starting structures with root mean square deviation to the native fold greater than 15 Å. Each replica run for 100 ns leading to an overall simulated time of 1 μs per protein. In Fig. S5 are shown two trajectories as well as a condensed sketch of all the terms of the potential.

VII. STRUCTURE-BASED REFERENCE SIMULATIONS

To set a reference baseline for the prediction quality of our coarse-grained model, we performed the same set of simulations for each of the 18 domains using the protein native contacts as the set of "predicted" contacts. These C_β-C_β native contacts have been obtained using a cutoff of 6.5 Å and excluding those contacts between first neighbor residues along the sequence. The number of native contact per protein is usually larger than N, the number of residues. The best dRMSD conformations obtained using this structure-based (SB) approach are shown as a black curve in Fig. S6.

VIII. PAS DOMAIN FOLDING

In the PAS domain we observe a large deviation between the dRMSD of the minimum energy structure (7 Å) and the absolute minimum dRMSD (3.6 Å) sampled during the simulation. Removing 5 predicted contacts belonging to the dimeric interface and repeating the simulation we observe a much better dRMSD of the minimum energy structure (4.7 Å) (see Fig. S6). To verify that this improvement is specific to these 5 contacts, we performed 20 independent simulations following the same protocol detailed above in the "Folding simulation details" paragraph, where in each of them, 5 randomly picked false positives contacts (CB-CB distance > 8 Å in the native structure) have been removed, paying attention to keep the 5 original dimeric contacts. None of these control simulations resulted in an improved dRMSD of the minimum energy structure.

IX. SRC CONFORMATIONAL SAMPLING

In Fig. S8 (panels A, B) we highlight the most flexible regions of SRC calculated using the root mean square fluctuation (RMSF). We observe a qualitative agreement with an extensive sampling all-atom molecular dynamics simulation of the Src domain [8] as shown in panel D. To be noted that in the all-atom simulation the transition of the A-loop had not been sampled, explaining in part the smaller fluctuations of such region in the all-atom RMSF. Furthermore, the temperature of the coarse-grained simulation is not directly comparable to the temperature of the all-atom simulation, preventing a sound quantitative comparison between the two RMSFs. To show if a residue flexibility is simply correlated to the number of its predicted interactions, we show in Fig. S8 (panels C) a representation where the color and thickness is inversely correlated to the number of contact predicted for each residue. Finally, in Fig. S8 (panel E) are shown the two sampled structures of SRC most similar to the active and inactive conformations respectively.

X. RAS FOLDING

We investigated the folding properties of the RAS domain (Pfam code PF00071, PDB 5P21, residues K5 to R164). To this aim, we used a 180 ns parallel tempering simulation with 10 replicas at increasing temperatures: 260.0 K, 270.0 K, 280.3 K, 291.0 K, 302.1 K, 313.7 K, 325.7 K, 338.2 K, 351.2 K, 364.7 K. We found that the folding transition of RAS features a clear peak in the heat capacity at constant volume $C_v(T)$ at T_f=317 K which defines the folding temperature in our model (see Fig. 9, top panel). Indeed, we find the transition to be highly cooperative and characterized by a van't Hoff enthalpy to calorimetric enthalpy ratio: κ_2=0.92. The Chan's parameter[9] κ_2 is defined as the ratio of the van't Hoff to calorimetric enthalpies: $\kappa_2 = 2T_{max}\sqrt{(k_B C_p(T_{max}))}/\Delta H_{cal}$, where T_{max} is the temperature of the C_p peak and ΔH_{cal} is the calorimetric enthalpy of the reaction determined as the integral of the heat capacity across the transition region.

In Fig. S9 (bottom panel) we show the free energy landscape close to T_f obtained using a standard C_α structure-based model. At variance with the model presented in this work, the absence of a folding intermediate in this case points to an interplay of the energetic and entropic contributions of the alpha helices to the stability of the protein that cannot be correctly captured with a too crude coarse-graining. The C_α structure-based simulation is done using the model of ref. [10] and [11] with default parameters.

XI. ANALYSIS DETAILS

If not otherwise stated, all the analysis are done on the lowest temperature trajectory. To compare the simulated protein X to the native conformation Y we use the distance root mean square deviation:

$$dRMSD = \sqrt{1/L(L-1)\sum_{i\neq j}[d(X_i,X_j)-d(Y_i,Y_j)]^2} \tag{18}$$

where $d(X_i, X_j)$ is the distance between C_α atom i and j of the structure X, calculated on the set of the L C_α atoms of the protein residues having less than 5% sequence gaps in the alignment. At variance with the commonly adopted RMSD, this measure does not need a previous roto-translation fit on a common structure to be meaningful.

The folding reaction cooperativity of Ras is estimated using the Chan's parameter[9] κ_2 defined as the ratio of the van't Hoff to calorimetric enthalpies: $\kappa_2 = 2T_{max}\sqrt{(}k_BC_p(T_{max}))/\Delta H_{cal}$, where T_{max} is the temperature of the C_p peak and ΔH_{cal} is the calorimetric enthalpy of the reaction determined as the integral of the heat capacity across the transition region.

To emulate a blind structure prediction where we don't have a reference structure to compare with, we report as our best predicted structures both the lowest energy structure sampled and the central structure of the most populated cluster. In both cases the lowest temperature trajectory is used. The central structure of the most populated cluster is obtained through a single-linkage clustering algorithm of g_cluster program from the GROMACS package with a C_α RMSD cutoff of 0.2 nm on 2500 structures uniformly picked in the second half of the simulation. The central structure of the cluster is the structure with the smallest distance to all of the other members of the cluster.

XII. DIHEDRAL POTENTIAL

The $V_{\phi,\psi}$ bonded term for GLY and non-GLY residues is obtained fitting the probability distribution of (ϕ, ψ) over the Ramachandran plot of two reference all-atom molecular dynamics simulations: one for a capped GLY peptide (ACE-GLY-NME) and one for a capped ALA peptide (ACE-ALA-NME). The water-solvated peptides have been prepared using a standard protocol: steepest descent minimization, 50 ps density equilibration at 300 K and constant pressure, 500 ps thermalization at constant volume and 2 ns equilibration at 300 K and constant volume with the velocity-rescale thermostat[12]. The AMBER99SB-ILDN force field [13] has been used for these simulations together with the TIP3P[16] water model for the 785 water molecules composing the systems. In order to have an exhaustive sampling of the $V_{ref}(\phi, \psi)$ potential energy surface (PES), we run the simulations for 20 ns, with a 2 ps time step, using the well-tempered metadynamics[17] enhanced sampling algorithm on the two dihedrals to guarantee a full convergence.

In the same way that the $V_{ref}(\phi, \psi)$ PES is the net result of the many contribution of the all-atom force field terms, we would like to reproduce this PES as the effective result of all the coarse-grained potential terms, including the yet to be determined $V_{\phi,\psi}$. In other words, to avoid double counting the interactions, $V_{\phi,\psi}$ should match $V_{diff}(\phi,\psi) = V_{ref}(\phi,\psi) - V_0(\phi,\psi)$ where $V_0(\phi,\psi)$ is the coarse-grained PES obtained without the (ϕ, ψ) dihedral terms.

To calculate $V_0(\phi, \psi)$ we simulated 20 ns of the corresponding coarse-grained peptides with the stripped down Hamiltonian: $H_0 = \widetilde{V}_{bonds} + \widetilde{V}_{angles} + \widetilde{V}_{torsions} + \widetilde{V}_{LJ}^{1-4}$ (see the main text for the terms explanation), using the the well-tempered metadynamics algorithm to have a reliable PES, as in the previous case.

Finally, in order to better fit the minima as compared to the transition barriers, the actual fit is done on the probability $p_{diff}(\phi,\psi) = exp(-V_{diff}(\phi,\psi)/kT)$ rather than directly on $V_{diff}(\phi,\psi)$. The fit of p_{diff} is done using the function $f(\phi,\psi) = exp(-[\sum_{i=1}^{5} k_i^{\phi}(1+cos(\phi-\phi_i^0)) + \sum_{i=1}^{5} k_i^{\psi}(1+cos(\psi-\psi_i^0))]/k_BT)$ which can be conveniently used in the MD code. The procedure is summed up in Fig. S10 together with the different PES involved. In particular, the resulting coarse-grained PES once all the terms are included is also shown as a double check. In table II, are reported the fit parameters.

[1] Lapedes A, Giraud B, Jarzynski C (2002) Using Sequence Alignments to Predict Protein Structure and Stability With High Accuracy. http://library.lanl.gov/cgi-bin/getfile?01038177.pdf.

[2] Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein–protein interaction by message passing. Proceedings of the National Academy of Sciences 106:67–72.

[3] Morcos F, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences 108:E1293–E1301.

[4] Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. Physical Review E 87:012707.

[5] Nesterov Y (2004) Introductory lectures on convex optimization (Springer Science & Business Media) Vol. 87.

[6] Dunn SD, Wahl LM, Gloor GB (2008) Mutual in-

formation without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24:333–340.

[7] Casari G, Sander C, Valencia A (1995) A method to predict functional residues in proteins. Nature structural biology pp 171–8.

[8] Lovera S, et al. (2012) The Different Flexibility of c-Src and c-Abl Kinases Regulates the Accessibility of a Druggable Inactive Conformation. J Am Chem Soc 134:2496–2499.

[9] Kaya H, Chan HS (2000) Polymer principles of protein calorimetric two-state cooperativity. Proteins 40:637–661.

[10] Clementi C, Nymeyer H, Onuchic JN (2000) Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. J Mol Biol 298:937–953.

[11] Noel JK, Whitford PC, Sanbonmatsu KY, Onuchic JN (2010) SMOG@ctbp: simplified deployment of structure-based models in GROMACS. Nucleic Acids Research 38:W657–61.

[12] Bussi G, Donadio D, Parrinello M (2007) Canonical sampling through velocity rescaling. J. Chem. Phys. 126:014101.

[13] Lindorff-Larsen K, et al. (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins 78:1950–1958.

[14] Cornell WD, et al. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 117:5179–5197.

[15] Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. J. Chem. Theory Comput. 4:435–447.

[16] Mahoney MW, Jorgensen WL (2000) A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. J. Chem. Phys. 112:8910.

[17] Barducci A, Bussi G, Parrinello M (2008) Well-tempered metadynamics: A smoothly converging and tunable free-energy method. Phys Rev Lett 100:20603–20604.

Pfam_ID	L	$\tau_{\rm slow}$	ϵ_r (10 $\tau_{\rm slow}$)
cNMP_binding	91	100	2.6
RRM_1	70	100	3.4
ADH_zinc_N	130	1000	3.5
TrkA_N	116	1000	2.9
CMD	85	800	1.5

Table I: The table shows the slowest relaxation time ($\tau_{\rm slow}$, in units of MC sweeps) associated to the MC dynamics in sequence space for five protein families (Pfam_ID) *of different length L, and the mean relative absolute error ϵ_r on frequencies reconstruction obtained using a number of MCMC steps proportional to $\tau_{\rm slow}$ in the evaluation of the gradient in Eq. 7.*

residue	i	k_i^ϕ	ϕ_i^0	k_i^ψ	ψ_i^0
GLY	1	-1.678	0.203	0.403	2.280
	2	13.973	6.130	9.963	4.389
	3	3.813	2.712	1.327	3.850
	4	-1.606	2.171	2.488	0.272
	5	-2.206	0.088	-1.121	6.167
non-GLY	1	9.668	5.396	-4.900	1.977
	2	9.194	3.662	5.627	2.217
	3	7.553	0.135	2.492	5.951
	4	-0.079	1.937	-2.265	3.372
	5	-0.527	2.162	0.635	4.165

Table II: The fit parameters of the $f(\phi,\psi)$ function for GLY and non-GLY. The units are kJ/mol for k_i^ϕ and k_i^ψ and radians for ϕ_i^0 and ψ_i^0.

PDB	N/Ng	prec	RMSD			dRMSD		
			$min_{Ng/N}$	CC	Emin	min_{Ng}	CC	Emin
1RQM	63/62	0.75	2.4/2.4	3.0	2.8	1.9	2.4	2.2
3F52	64/49	0.69	1.7/3.0	2.4	2.3	1.3	1.9	1.9
1OR7	68/53	0.74	1.6/2.4	2.3	2.9	1.0	1.3	1.9
1G2E	71/50	0.80	1.6/2.3	2.4	2.5	1.2	2.0	2.0
1ODD	76/65	0.55	3.0/3.5	5.0	4.0	2.1	3.4	3.1
3FHI	81/72	0.89	2.1/2.7	2.6	2.9	1.6	1.9	2.1
3D7I	85/61	0.39	3.9/6.4	6.2	6.4	2.4	3.7	3.8
3DF8	87/77	0.48	3.0/3.6	4.0	3.7	2.2	2.8	2.6
1BQU	88/57	0.58	2.7/4.0	4.1	4.7	2.1	2.8	2.9
2O72	90/66	0.69	3.4/3.7	3.9	4.1	2.5	3.0	3.1
1OAP	96/78	0.63	2.2/3.3	3.0	3.2	1.8	2.3	2.4
1KGS	111/99	0.70	3.0/3.1	3.6	4.2	2.5	2.9	3.1
2GJ3	112/80	0.58	4.8/5.3	11.8	10.0	3.6	9.8	7.3
3NYY	112/82	0.71	3.7/6.1	5.0	4.8	2.7	3.2	3.5
3FWZ	116/100	0.73	2.9/3.5	3.7	4.5	2.1	2.6	3.0
1A71	119/99	0.66	4.6/5.5	6.1	6.2	2.9	3.7	3.6
5P21	160/144	0.73	3.6/3.7	4.3	4.4	2.7	3.3	3.3
3TGI	216/167	0.78	3.8/4.2	5.3	5.8	2.8	3.7	4.0

Table III: The set of 18 proteins simulated is shown with their PDB code together with the total number of residues N and the number of residues with less than 5% gaps Ng as a subscript, the precision for the top N predictions prec, the minimum root mean square deviation (RMSD) structure calculated over the Ng residues and over all N residues $min_{Ng/N}$, the RMSD of the central cluster structure CC and the RMSD of the minimum energy structure Emin. In the last three columns are shown the minimum distance RMSD (dRMSD) structure calculated over the Ng residues min_{Ng} and the dRMSD of the central cluster structure CC as well as the dRMSD of the minimum energy structure Emin. Both the RMSD and dRMSD are calculated using the CA atoms and the values are in Angstroms (Å).

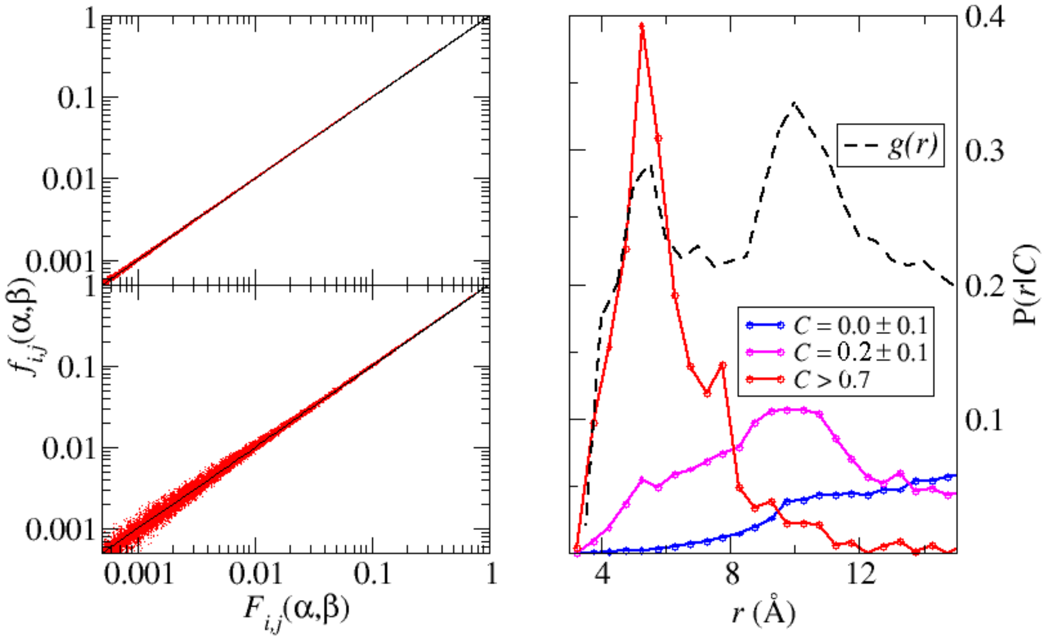

Figure 1: **Left panel***: correlation between the pairwise frequencies $f_{i,j}(\alpha,\beta)$ (plotted on y-axes) computed from Metropolis Monte Carlo sampling of the pairwise model and the empirical frequencies $F_{i,j}(\alpha,\beta)$ from multiple sequence alignment (x-axes) (after correcting for the bias from regularization, see Eq. 7); on top, results are shown for one of the smallest domains (RRM_1 domain, 71 residues); on bottom, for the largest protein in the dataset (Trypsin domain, 216 residues).* **Right panel***: the residue-residue, C_β-C_β distance distributions for pairs corresponding to different values of coevolutionary coupling C are shown as continuous lines (large couplings, $C > 0.7$, red line; weak couplings, $C = 0.2 \pm 0.1$, magenta line; couplings close to zero, $C = 0.0 \pm 0.1$, blue line). The black broken line corresponds to the distribution for the residue-residue distances, computed from the 18 experimental structures and normalized for the ideal gas contribution ($g(r) \propto P(r)/r^2$) (arbitrary scale).*

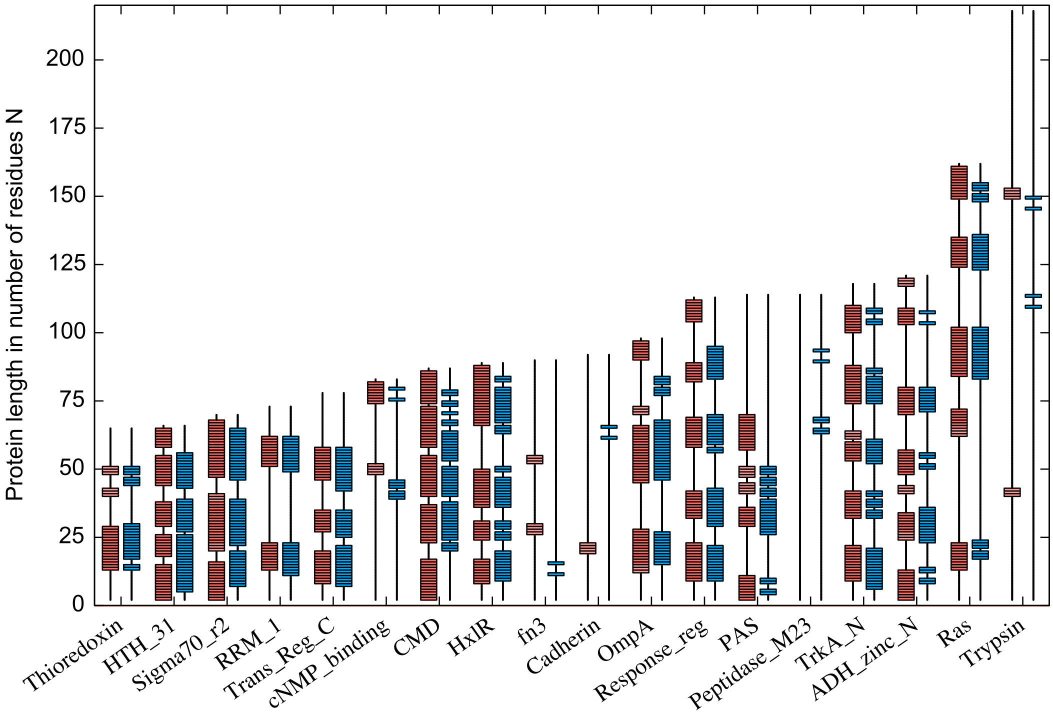

Figure 2: Schematic comparison of the actual (red, pink) and the predicted (cyan) alpha helical content per residue of the 18 proteins. In pink are shown the 3-10 helices.

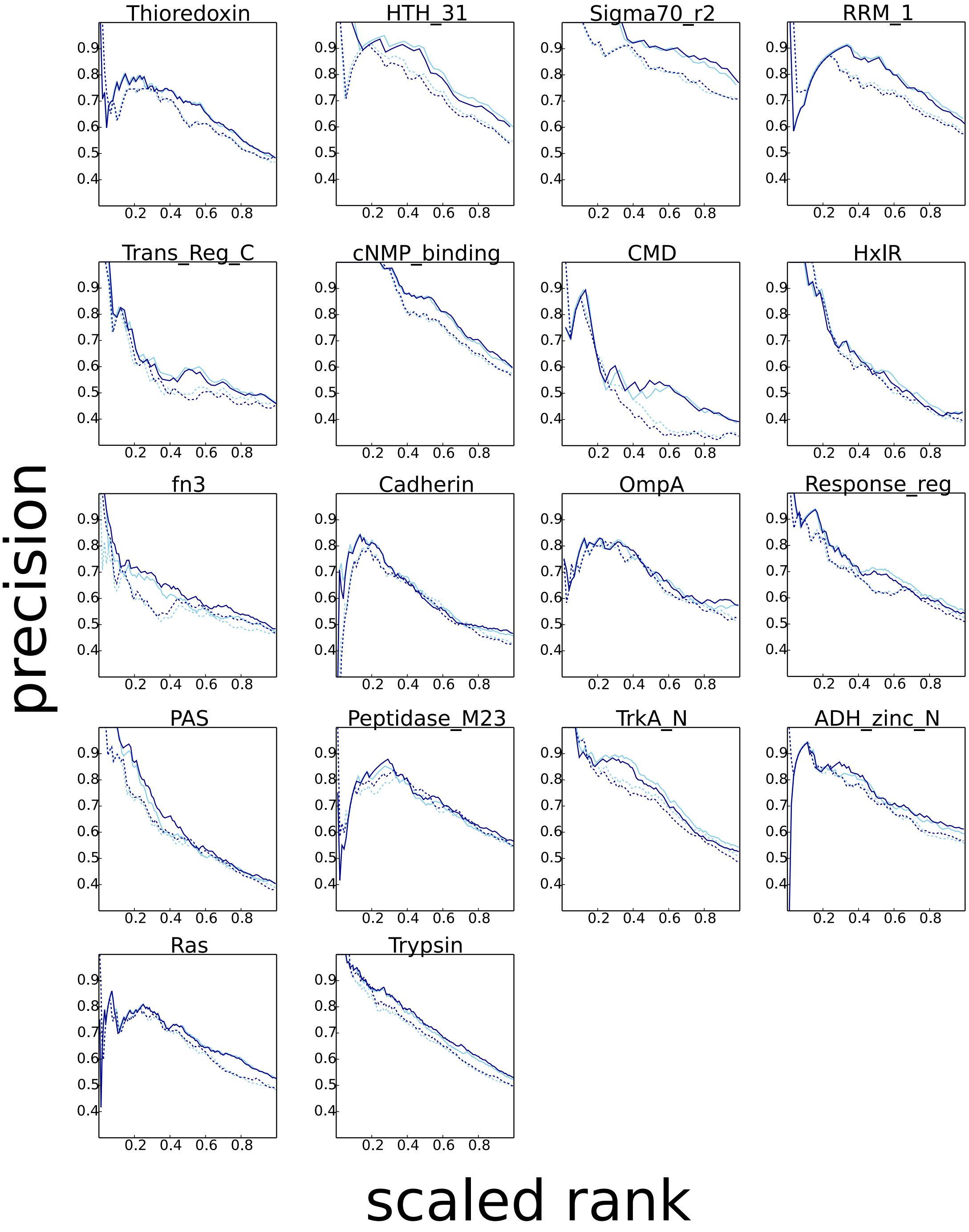

Figure 3: Precision of residue-residue contact prediction for the 18 protein domains, as a function of the scaled rank of the score, defined as rank / total number of contacts in a structure. Continuous lines show results for the Boltzmann learning algorithm, while the broken lines refer to the mean field solution for the couplings[3]. We included in the analysis all the residue pairs with a sequence separation larger than four (dark blue lines) and larger than five (light blue lines).

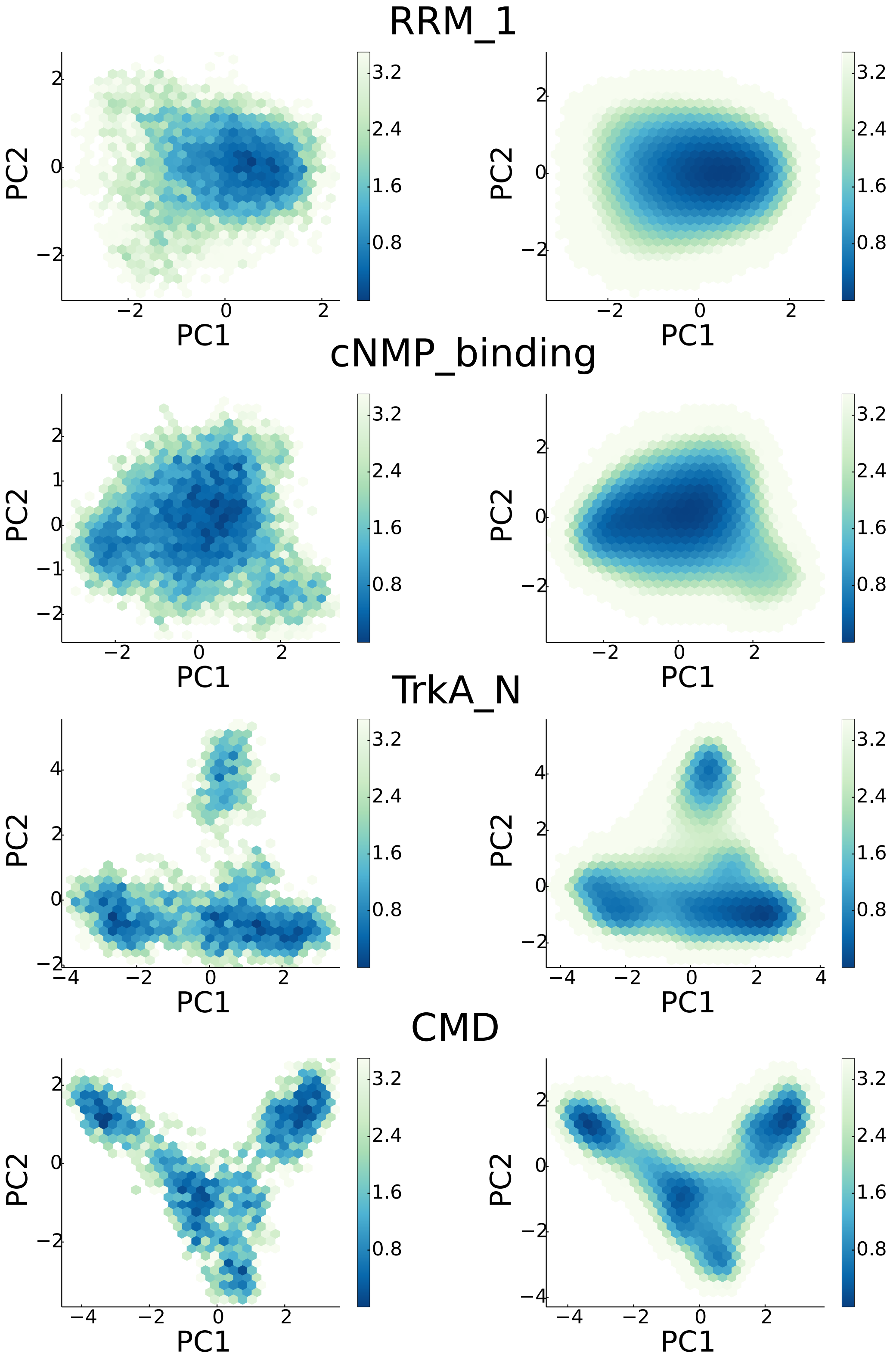

Figure 4: "energy" landscapes in the space of the first two principal components of the MSA covariance matrix for four protein families: (from top to bottom) RRM_1, cNMP_binding, TrkA_N and CMD.

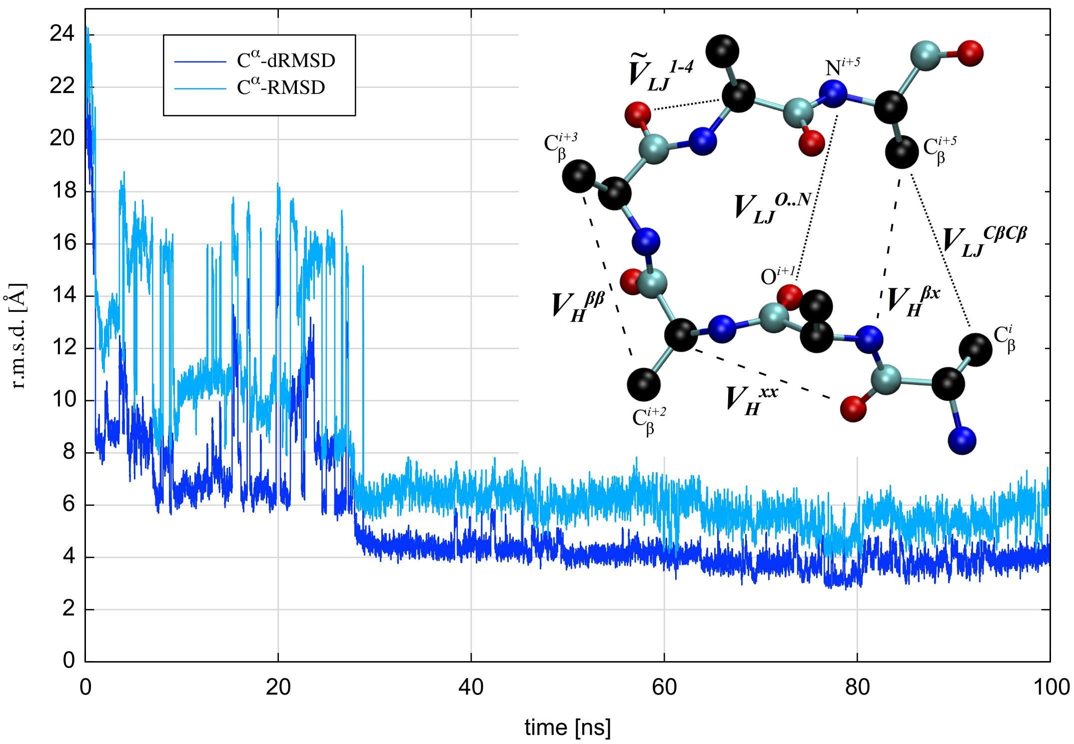

Figure 5: The dRMSD and RMSD calculated over the lowest temperature trajectory for the largest protein, Trypsin. In the inset, a short protein fragment to illustrate the level of coarse-graining and all the non-bonded interactions present in the coarse-grained model. The dashed lines indicate a Lennard Jones potential, the dotted lines a repulsive potential. See the Model description for their functional form and parameters. In red are shown the O atoms, in black the C_α and C_β atoms, in blue the N atoms and in gray the C' atoms.

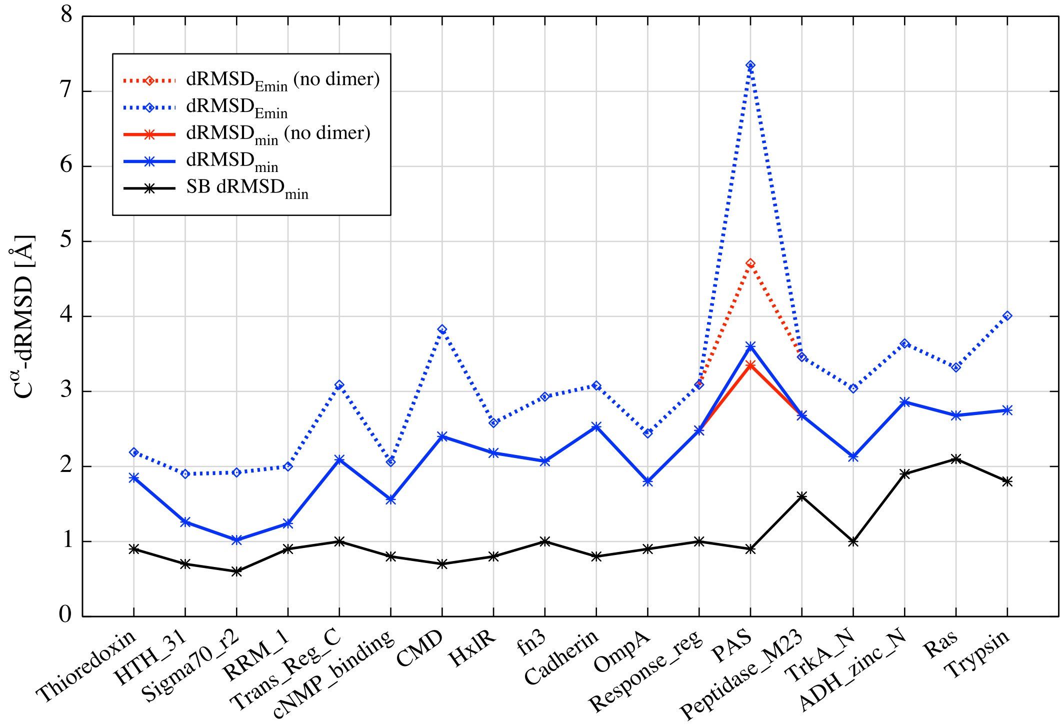

Figure 6: The C_α-dRMSD of the predicted structure with respect to the native conformation calculated using all the residues with less than 5% gaps in the sequence alignment. The proteins, shown with their Pfam ID, are listed with increasing length from left to right. The continuous blue line connects the best structure dRMSD sampled during the run while the dashed blue line the dRMSD of the minimum energy structure sampled. Also shown as a baseline (continuous black curve) are the best dRMSD obtained using the same protein model where instead of the predicted contacts the actual native contacts are used as in a structure-based (SB) model. The lines are merely to guide the eye.

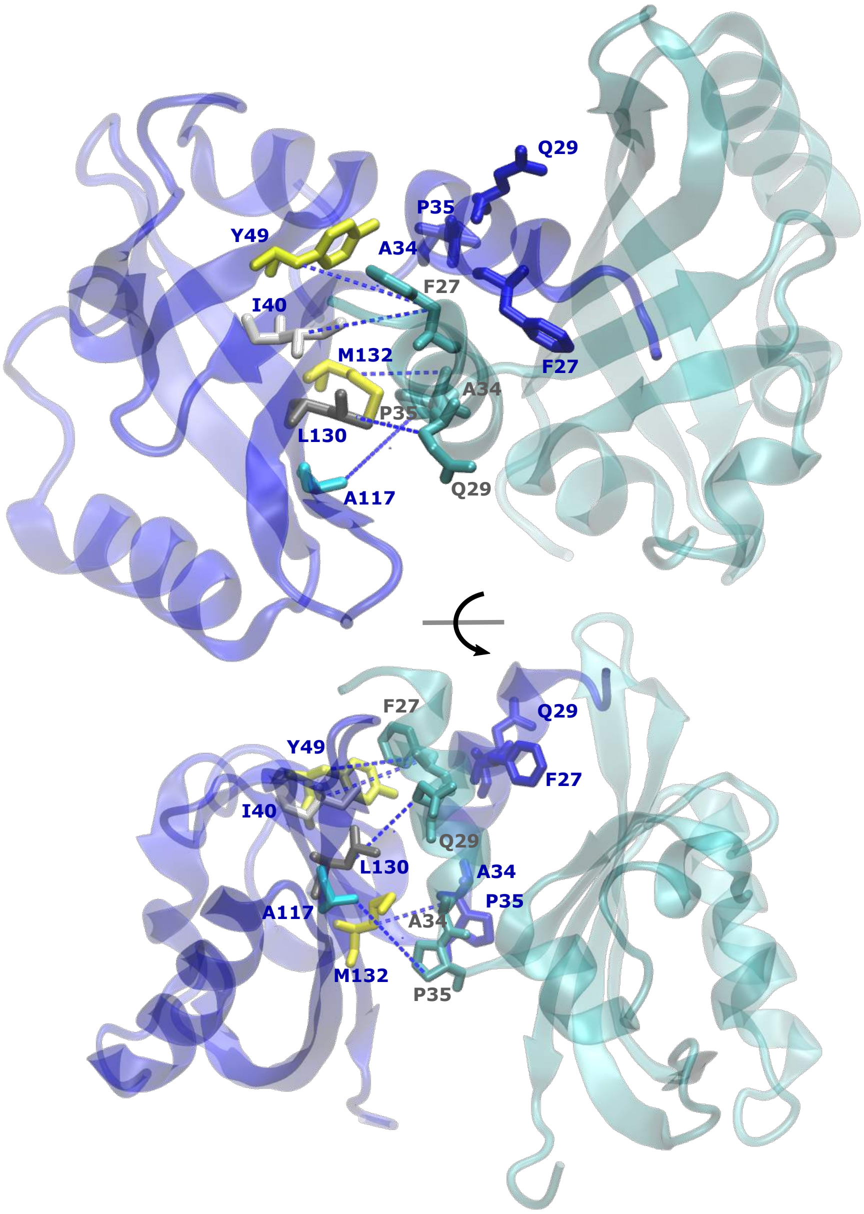

Figure 7: The chain A and B of PAS homodimer are shown in blue and cyan respectively, together with the 5 dimer interface contacts: I40 - F27 ; Y49 - F27 ; A117 - P35 ; L130 Q29 ; M132 - A34 (residue chain A - residue chain B).

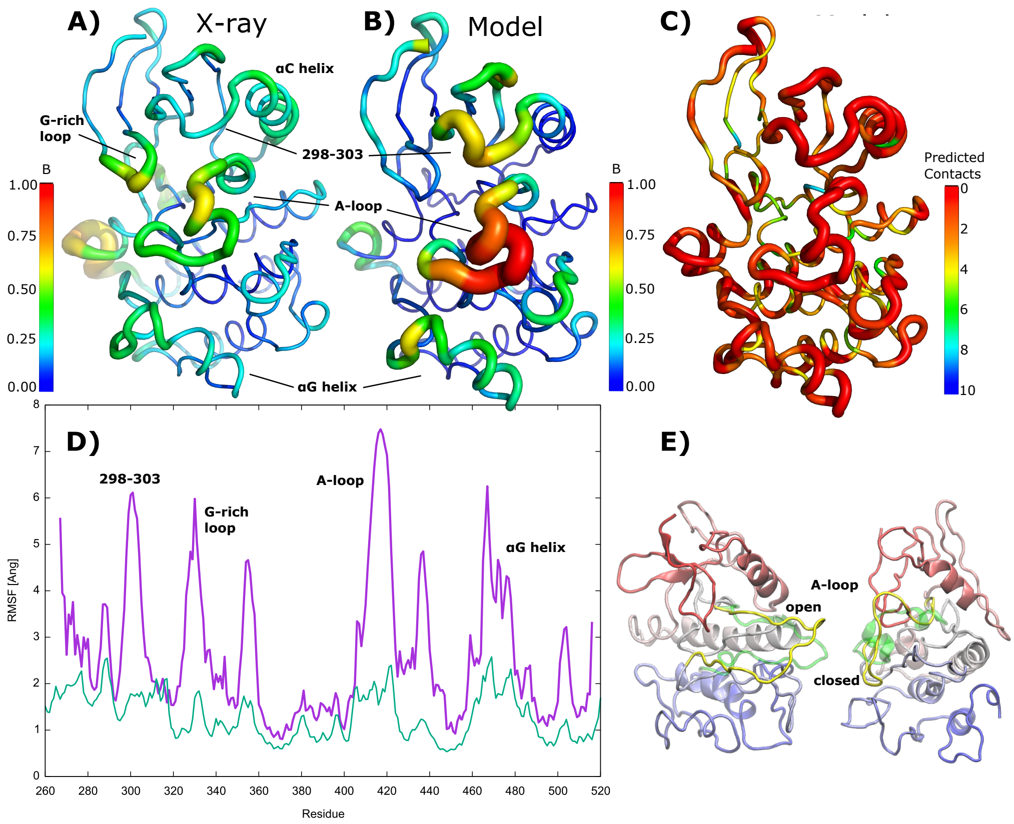

Figure 8: The B-factor of the X-ray 2SRC structure (panel A) is compared to the same quantity obtained during the simulation (panel B). Red zones represent more flexible regions and the thickness is proportional to the value. In panel C we report on the structure the number of predicted contacts per each residue. No contacts predicted correspond to red color. In panel D we show the root mean square fluctuation (RMSF) of SRC catalytic domain obtained with the model simulations (blue) overlaid on the RMSF obtained using a reference all-atom MD simulation (green). The RMSF are translated on the structure as B-factors using the relation $B = (8\pi^2/3)RMSF^2$*. In panel E we show two sampled conformations of the SRC catalytic domain most similar to the active state with an open activation loop (left) and inactive state with a closed loop (right). The activation loop (A-loop) is shown in yellow in the sampled conformations and in transparent green for the active (PDB: 1Y57) and inactive (PDB: 2SRC) crystal structures.*

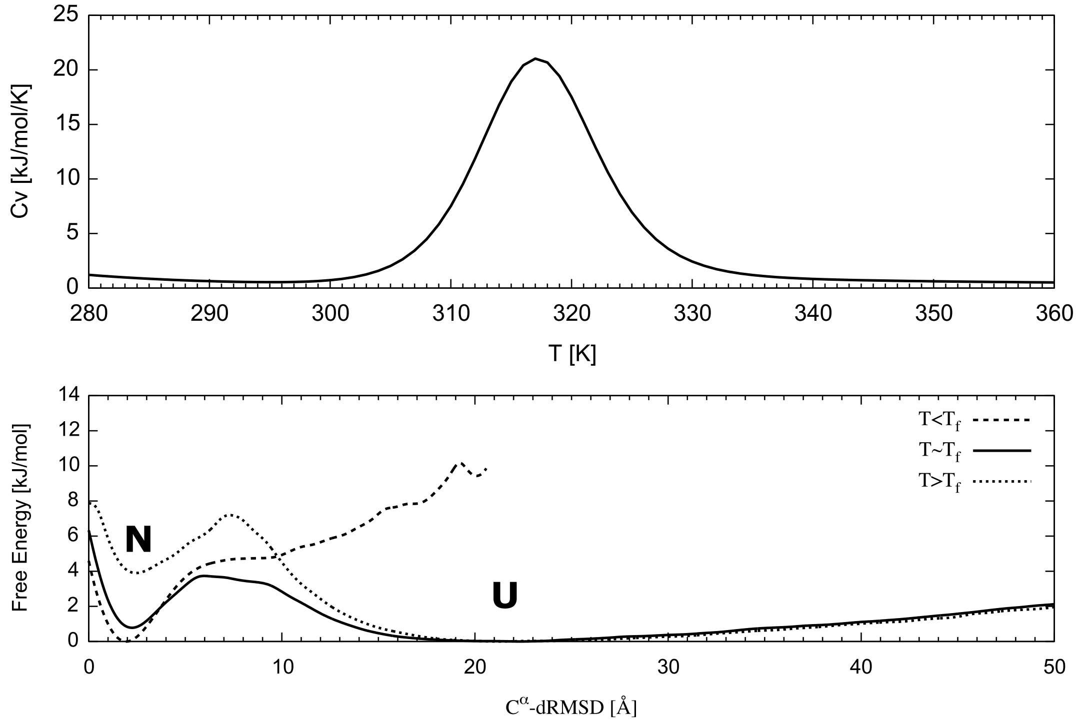

Figure 9: Thermodynamics of RAS folding. Top panel: the heat capacity C_v is reported as a function of temperature T obtained simulating RAS with our model. Bottom panel: the folding free energy of Ras protein using a C_α structure-based model ([10, 11]) shown at three different temperatures around the folding temperature T_f as a function of the C_α-dRMSD. The native (N) and unfolded (U) states are also identified.

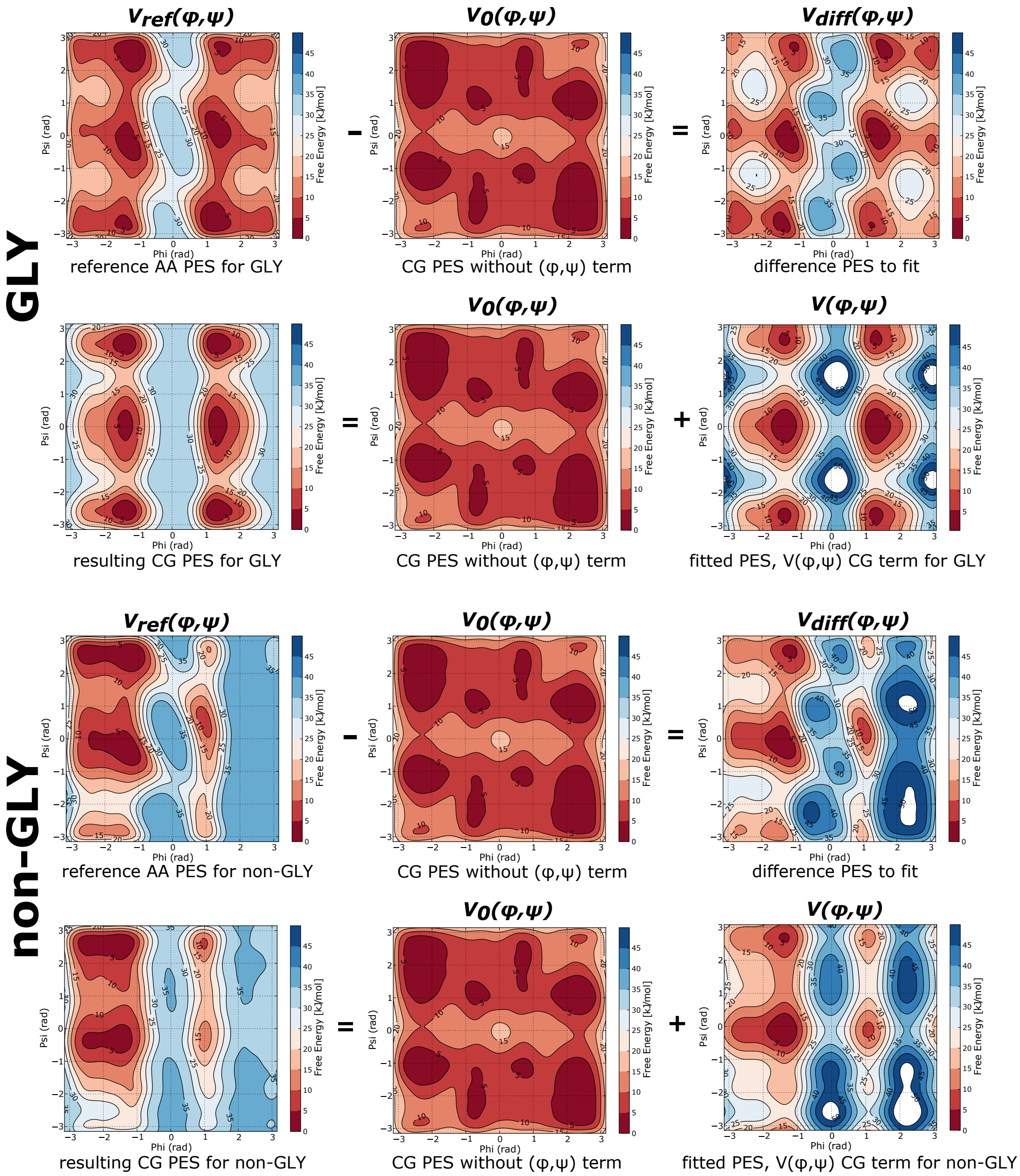

Figure 10: The protocol to obtain the coarse-grained dihedral term $V(\phi,\psi)$ is represented schematically as two equations (top two rows for GLY, bottom two rows for non-GLY). The $V_{ref}(\phi,\psi)$ is the desired potential energy surface obtained through an all-atom (AA) simulation. Subtracting the PES that results from the coarse-grained (CG) model without any (ϕ,ψ) term ($V_0(\phi,\psi)$, central column) we obtain the PES to be fitted $V_{diff}(\phi,\psi)$. Using the fit described in the text, we obtain the CG potential term $V(\phi,\psi)$ for each residue (second row for glycine, bottom row for non-GLY, on the right hand side). As a check, if we now simulate the capped glycine peptide (or alanine peptide) with all the CG terms, we obtain the PES shown on the left hand side, second row (bottom row for alanine), for which the shape and position of the main minima reproduces the AA one ($V_{ref}(\phi,\psi)$).